TMA4267 Linear statistical models

11. march 2025

Thea Bjørnland

TMA4267 Linear statistical models

About the course

Examination arrangement

Examination arrangement: School exam
Grade: Letter grades

Evaluation	Weighting	Duration	Examination aids
School exam	100/100	4 hours	C

Course content

~~Random vectors. Multivariate normal distribution. Multiple linear regression.~~ Analysis of variance. ~~Multiple hypothesis testing~~. Design of experiments.

Analysis of variance (ANOVA)

Härdle and Simar, chapter 8.1.1

8.1.1 ANOVA Models

One-Factor Models

In Sect. 3.5, we introduced the example of analysing the effect of one factor (three possible marketing strategies) on the sales of a product (a pullover), see Table 3.2. The standard way to present one factor ANOVA models with p levels is as follows

$$y_{k\ell} = \mu + \alpha_\ell + \varepsilon_{k\ell}, \; k = 1, \ldots, n_\ell, \text{ and } \ell = 1, \ldots, p, \tag{8.2}$$

all the $\varepsilon_{k\ell}$ being independent. Here ℓ is the label which indicates the level of the factor and α_ℓ is the effect of the ℓth level: it measures the deviation from μ, the global mean of y, due to this level of the factor. In this notation, we need to impose the restriction $\sum_{\ell=1}^{p} \alpha_\ell = 0$ in order to identify μ as the mean of y. This presentation is equivalent, but slightly different, to the one presented in Chap. 3 (compare with Eq. (3.41)), but it allows for easier extension to the multiple factors case. Note also that here we allow different sample sizes for each level of the factor (an unbalanced design, more general than the balanced design presented in Chap. 3).

Recall from TMA4240/45

We have two populations with means μ_1 and μ_2.

We have (independent) random samples of sizes n_1 and n_2 from each of the two populations:

$$Y_{1,1}, Y_{1,2}, \ldots, Y_{1,n_1} \qquad Y_{1,j} \sim N(\mu_1, \sigma^2),\ \ j = 1, \ldots, n_1$$

$$Y_{2,1}, Y_{2,2}, \ldots, Y_{2,n_2} \qquad Y_{2,j} \sim N(\mu_2, \sigma^2),\ \ j = 1, \ldots, n_2$$

We test $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 \neq \mu_2$ with the t-test:

$$T = \frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{s_p^2(1/n_1 + 1/n_2)}}$$

What if we have many groups?

Recall from earlier this semester

$y = X\beta + \varepsilon$

n subjects

$p + 1$ regression coefficients

Partitioning of variation (sums of squares)

$SS_T = SS_R + SS_E$

$SS_T = \sum_{i=1}^{n} (y_i - \bar{y})^2$ $SS_R = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$ $SS_E = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$

$SS_T = y^T(I - \frac{1}{n}J)y$ $SS_R = y^T(H - \frac{1}{n}J)y$ $SS_E = y^T(I - H)y$

Recall from earlier this semester

$$y = X\beta + \varepsilon$$

n subjects

$p+1$ regression coefficients

Analysis of variance table for table for linear regression

Source	Sum of squares	DF	Mean sum of squares	F
Regression	$SS_R = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$	p	SS_R/p	$\frac{SS_R/p}{SS_E/(n-(p+1))}$
Error	$SS_E = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$	$n-(p+1)$	$SS_E/(n-(p+1))$	
Total	$SS_T = \sum_{i=1}^{n} (y_i - \bar{y})^2$	$n-1$		

3.13 Testing Linear Hypotheses

Hypotheses

1. General linear hypothesis:

$$H_0 : \boldsymbol{C}\boldsymbol{\beta} = \boldsymbol{d} \qquad \text{against} \qquad H_0 : \boldsymbol{C}\boldsymbol{\beta} \neq \boldsymbol{d}$$

   where $\boldsymbol{C}$ is a $r \times p$-matrix with $\text{rk}(\boldsymbol{C}) = r \leq p$ (r linear independent restrictions).

2. Test of significance (t-test):

$$H_0 : \beta_j = 0 \qquad \text{against} \qquad H_1 : \beta_j \neq 0$$

3. Composite test of a subvector:

$$H_0 : \boldsymbol{\beta}_1 = \boldsymbol{0} \qquad \text{against} \qquad H_1 : \boldsymbol{\beta}_1 \neq \boldsymbol{0}$$

4. Test for significance of regression:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0 \text{ against}$$
$$H_1 : \beta_j \neq 0 \text{ for at least one } j \in \{1, \ldots, k\}$$

Test Statistics

Assuming normal errors we obtain under H_0:

1. $F = 1/r\,(\boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{d})' \left(\hat{\sigma}^2 \boldsymbol{C}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{C}'\right)^{-1} (\boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{d}) \sim \text{F}_{r,n-p}$
2. $t_j = \dfrac{\hat{\beta}_j}{\text{se}_j} \sim \text{t}_{n-p}$
3. $F = \frac{1}{r}(\hat{\boldsymbol{\beta}}_1)'\widehat{\text{Cov}(\hat{\boldsymbol{\beta}}_1)}^{-1}(\hat{\boldsymbol{\beta}}_1) \sim \text{F}_{r,n-p}$
4. $F = \dfrac{n-p}{k}\dfrac{R^2}{1-R^2} \sim \text{F}_{k,n-p}$

Analysis of variance (ANOVA)

Analysis of variance

46 languages

Article Talk Read Edit View history Tools

From Wikipedia, the free encyclopedia

Analysis of variance (ANOVA) is a family of statistical methods used to compare the means of two or more groups by analyzing variance. Specifically, ANOVA compares the amount of variation *between* the group means to the amount of variation *within* each group. If the between-group variation is substantially larger than the within-group variation, it suggests that the group means are likely different. This comparison is done using an F-test. The underlying principle of ANOVA is based on the law of total variance, which states that the total variance in a dataset can be broken down into components attributable to different sources. In the case of ANOVA, these sources are the variation between groups and the variation within groups.

ANOVA was developed by the statistician Ronald Fisher. In its simplest form, it provides a statistical test of whether two or more population means are equal, and therefore generalizes the *t*-test beyond two means.

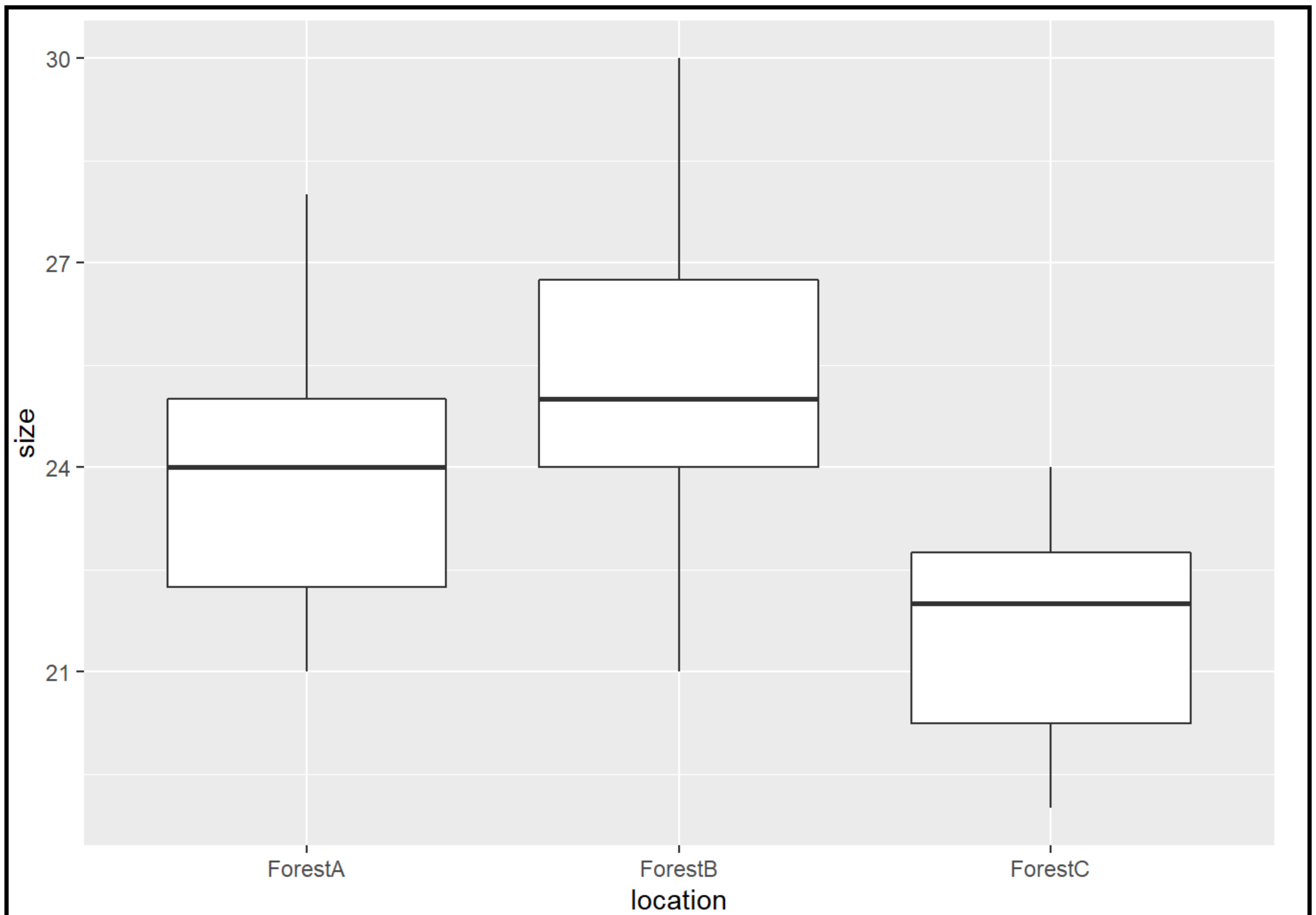
Generalizations [edit]

ANOVA is considered to be a special case of linear regression[59][60] which in turn is a special case of the general linear model.[61] All consider the observations to be the sum of a model (fit) and a residual (error) to be minimized.

https://en.wikipedia.org/wiki/Analysis_of_variance *https://biostats.w.uib.no/one-way-anova/*

One-factor anova

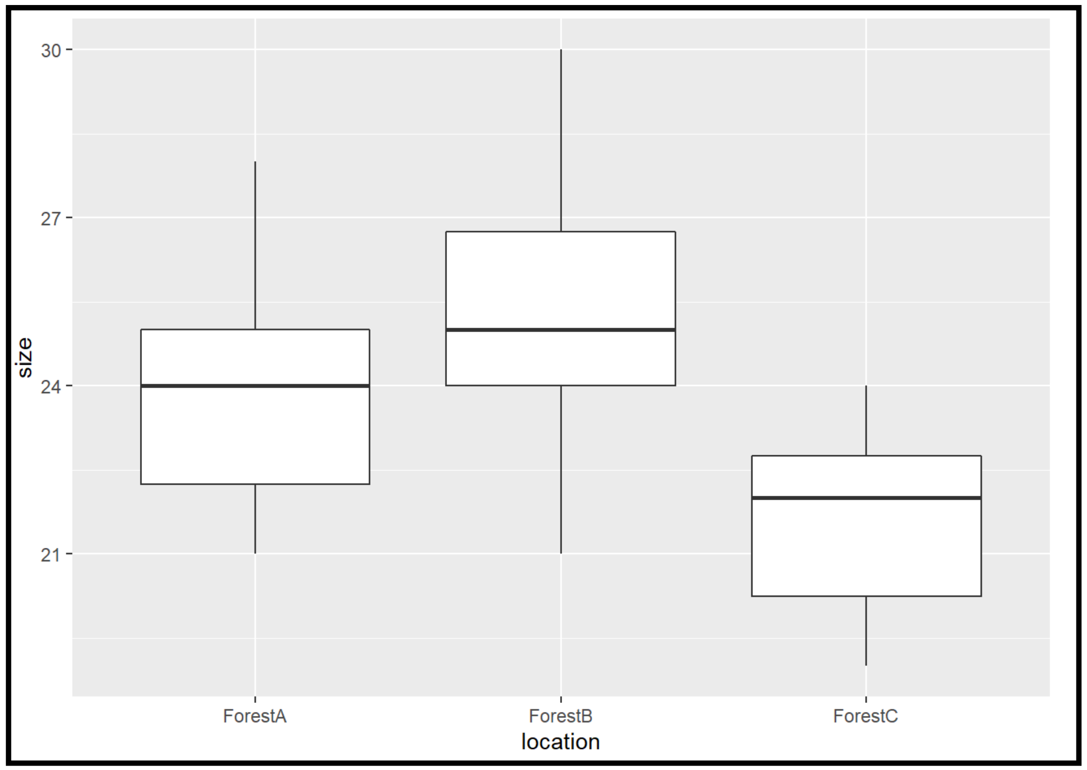
Example: We want to check whether the average size of blue ground beetles (*Carabus intricatus*) differs depending on their location. We consider 3 different locations, A, B and C, and we measure the size (in millimeters) of 10 individuals at each location.



Are the beetles different between locations, or is this something we can expect to occur "by chance"?

https://biostats.w.uib.no/one-way-anova/

One-factor anova



$$\text{Between-group variability} = \sum_{j=1}^{3}\sum_{i=1}^{10}(\bar{y}_j - \bar{y})^2$$

Discuss (2-3 min):

Why is this expression useful for testing if all group means are equal?

Try to make a sketch of the distribution under the null hypothesis.

https://biostats.w.uib.no/one-way-anova/

One-factor anova

```
Call:
lm(formula = size ~ location, data = beetleds)

Residuals:
   Min     1Q Median     3Q    Max 
-4.300 -1.300 -0.150  1.375  4.700 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)      24.0000     0.6966  34.454   <2e-16 ***
locationForestB   1.3000     0.9851   1.320      0.2    
locationForestC  -1.5000     1.1375  -1.319      0.2    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.203 on 23 degrees of freedom
Multiple R-squared:  0.2107,	Adjusted R-squared:  0.142 
F-statistic: 3.069 on 2 and 23 DF,  p-value: 0.06584
```

Dummy coding

https://biostats.w.uib.no/one-way-anova/

One-factor anova

```
Call:
lm(formula = size ~ location, data = beetleds, contrasts = list(location = "contr.sum"))

Residuals:
   Min     1Q Median     3Q    Max 
-4.300 -1.300 -0.150  1.375  4.700 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 23.93333    0.44461  53.830   <2e-16 ***
location1    0.06667    0.59952   0.111   0.9124    
location2    1.36667    0.59952   2.280   0.0322 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.203 on 23 degrees of freedom
Multiple R-squared:  0.2107,	Adjusted R-squared:  0.142 
F-statistic: 3.069 on 2 and 23 DF,  p-value: 0.06584
```

Effect coding

https://biostats.w.uib.no/one-way-anova/

One-factor anova

In R: anova(model)

```
Analysis of Variance Table

Response: size
          Df  Sum Sq Mean Sq F value  Pr(>F)
location   2  29.785 14.8923  3.0692 0.06584 .
Residuals 23 111.600  4.8522
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Summing up: One-factor anova

$$y_{ij} = \mu + \alpha_j + \varepsilon_{ij} \qquad \varepsilon_{ij} \sim N(0, \sigma^2) \qquad \sum_j \alpha_j = 0$$

Aim: estimate parameters of the model and test:

$H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_m = 0$

H_1 : at least one α_j differs

$y = X\beta + \varepsilon$ "Effect coding" of design matrix

Gives estimates of μ and $m - 1$ effects (the *m*-th effect found implicitly)

Test H_0 vs H_1 by testing significance of regression

Two-factor anova

Machine operator example

```
> head(ds,8)
     time operator machine
[1,] 42.5        1       1
[2,] 39.8        1       2
[3,] 40.2        1       3
[4,] 41.3        1       4
[5,] 39.3        2       1
[6,] 40.1        2       2
[7,] 40.5        2       3
[8,] 42.2        2       4
```

Two-factor anova

Machine operator example

	(Intercept)	machine1	machine2	machine3	operator1	operator2	operator3	operator4	operator5
1	1	1	0	0	1	0	0	0	0
2	1	0	1	0	1	0	0	0	0
3	1	0	0	1	1	0	0	0	0
4	1	-1	-1	-1	1	0	0	0	0
5	1	1	0	0	0	1	0	0	0
6	1	0	1	0	0	1	0	0	0
7	1	0	0	1	0	1	0	0	0
8	1	-1	-1	-1	0	1	0	0	0
9	1	1	0	0	0	0	1	0	0
10	1	0	1	0	0	0	1	0	0
11	1	0	0	1	0	0	1	0	0
12	1	-1	-1	-1	0	0	1	0	0
13	1	1	0	0	0	0	0	1	0
14	1	0	1	0	0	0	0	1	0
15	1	0	0	1	0	0	0	1	0
16	1	-1	-1	-1	0	0	0	1	0
17	1	1	0	0	0	0	0	0	1
18	1	0	1	0	0	0	0	0	1
19	1	0	0	1	0	0	0	0	1
20	1	-1	-1	-1	0	0	0	0	1
21	1	1	0	0	-1	-1	-1	-1	-1
22	1	0	1	0	-1	-1	-1	-1	-1
23	1	0	0	1	-1	-1	-1	-1	-1
24	1	-1	-1	-1	-1	-1	-1	-1	-1

Two-factor anova

Machine operator example

```
Call:
lm(formula = time ~ machine + operator, data = ds, contrasts = list(machine = "contr.sum",
    operator = "contr.sum"))

Residuals:
    Min      1Q  Median      3Q     Max 
-2.3375 -0.5437  0.1625  0.6000  2.3708 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  42.1208     0.2574 163.653  < 2e-16 ***
machine1     -0.8208     0.4458  -1.841  0.08544 .  
machine2     -0.7375     0.4458  -1.654  0.11882    
machine3      0.4458     0.4458   1.000  0.33313    
operator1    -1.1708     0.5755  -2.034  0.05999 .  
operator2    -1.5958     0.5755  -2.773  0.01422 *  
operator3    -0.8958     0.5755  -1.557  0.14042    
operator4     0.3292     0.5755   0.572  0.57583    
operator5     1.9292     0.5755   3.352  0.00437 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.261 on 15 degrees of freedom
Multiple R-squared:  0.7087,	Adjusted R-squared:  0.5533 
F-statistic: 4.561 on 8 and 15 DF,  p-value: 0.005599
```

3.13 Testing Linear Hypotheses

Hypotheses

1. General linear hypothesis:

$$H_0 : \boldsymbol{C}\boldsymbol{\beta} = \boldsymbol{d} \qquad \text{against} \qquad H_0 : \boldsymbol{C}\boldsymbol{\beta} \neq \boldsymbol{d}$$

   where $\boldsymbol{C}$ is a $r \times p$-matrix with $\text{rk}(\boldsymbol{C}) = r \leq p$ (r linear independent restrictions).

2. Test of significance (t-test):

$$H_0 : \beta_j = 0 \qquad \text{against} \qquad H_1 : \beta_j \neq 0$$

3. Composite test of a subvector:

$$H_0 : \boldsymbol{\beta}_1 = \boldsymbol{0} \qquad \text{against} \qquad H_1 : \boldsymbol{\beta}_1 \neq \boldsymbol{0}$$

4. Test for significance of regression:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0 \text{ against}$$
$$H_1 : \beta_j \neq 0 \text{ for at least one } j \in \{1, \ldots, k\}$$

Test Statistics

Assuming normal errors we obtain under H_0:

1. $F = 1/r\,(\boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{d})' \left(\hat{\sigma}^2 \boldsymbol{C}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{C}'\right)^{-1} (\boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{d}) \sim \text{F}_{r,n-p}$
2. $t_j = \dfrac{\hat{\beta}_j}{\text{se}_j} \sim \text{t}_{n-p}$
3. $F = \frac{1}{r}(\hat{\boldsymbol{\beta}}_1)'\widehat{\text{Cov}(\hat{\boldsymbol{\beta}}_1)}^{-1}(\hat{\boldsymbol{\beta}}_1) \sim \text{F}_{r,n-p}$
4. $F = \dfrac{n-p}{k}\,\dfrac{R^2}{1-R^2} \sim \text{F}_{k,n-p}$

Two-factor anova

Machine operator example

In R: anova(model)

```
Analysis of Variance Table

Response: time
          Df Sum Sq Mean Sq F value   Pr(>F)
machine    3 15.925  5.3082  3.3388 0.047904 *
operator   5 42.087  8.4174  5.2944 0.005328 **
Residuals 15 23.848  1.5899
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```